SLoRA: Learning Structured Low-Rank Updates for Cross-Domain Transfer

Anonymous authors
Paper under double-blind review

Abstract

Parameter-efficient fine-tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) have become standard tools for adapting large pre-trained models. However, conventional LoRA applies uniform low-rank updates across all adaptation directions, which can limit flexibility and lead to unstable optimization or degraded generalization under domain shift, particularly in multimodal settings. We propose SLoRA, a lightweight extension of LoRA that introduces structured basis-space reparameterization and adaptive gating into low-rank updates to improve stability and robustness during transfer. Specifically, SLoRA transforms the low-rank adaptation space into a fixed orthogonal basis representation, applies learnable basis scaling to redistribute adaptation capacity across structured directions, and employs local basis modulation together with a global input-adaptive gate to control update strength. These mechanisms are batch-stabilized, allowing SLoRA to operate reliably on mixed-modality data without requiring explicit modality labels. We further provide theoretical insights suggesting that the proposed design promotes smoother update dynamics and acts as a form of structured adaptation regularization. Empirically, we evaluate SLoRA across four diverse domains, radiology, pathology, remote sensing, and art, on six multimodal visual question answering (VQA) benchmarks and eleven image classification datasets using multiple general-purpose vision and vision-language backbones. SLoRA consistently outperforms standard LoRA, achieving average gains of over 2.2% on VQA tasks and 2.9% on classification tasks, while remaining fully plug-and-play and parameter-efficient. These results demonstrate that structured low-rank reparameterization with input-adaptive modulation provides an effective and practical solution for robust parameter-efficient transfer under domain shift.

1 Introduction

Large multimodal foundation models such as CLIP Radford et al. (2021), BLIP Li et al. (2022b), and GIT Wang et al. (2022) have demonstrated impressive generalization across a wide range of tasks. However, when transferred to specialized downstream domains, such as medical imaging, art, or remote sensing, their performance often degrades substantially due to domain shifts in data distribution, visual characteristics, and semantic alignment. While full fine-tuning can mitigate this issue, it requires updating all backbone parameters, resulting in prohibitive computational costs that limit its practicality in real-world adaptation scenarios.

Parameter-efficient fine-tuning (PEFT) Houlsby et al. (2019); Ding et al. (2023) has therefore emerged as a practical alternative, achieving competitive performance while updating only a small fraction of model parameters. Among them, LoRA Hu et al. (2022) injects low-rank trainable modules into linear layers, enabling lightweight and plug-and-play adaptation. Despite its efficiency, LoRA applies uniform adaptation across all low-rank directions, without explicitly modeling their different contributions under domain shift. Consequently, its adaptation capacity may be distributed suboptimally, leading to unstable optimization and reduced robustness when transferring across domains. Recent studies have explored more structured parameterization strategies for adaptation, including methods based on structured basis representations or transformed adaptation spaces Zhang & Pilanci (2024); Borse et al. (2024); Liu et al. (2024b). While these approaches demonstrate the benefits of structured reparameterization, they are often designed for specific architectures or rely on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

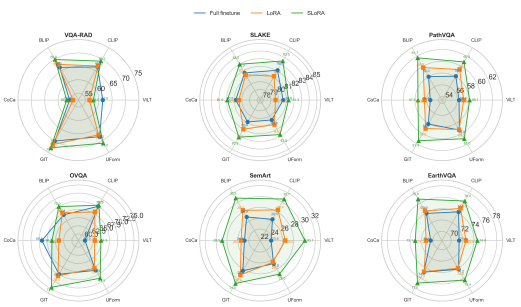

Figure 1: Performance of full fine-tuning, LoRA, and SLoRA across six domain-specific VQA datasets and multiple models.

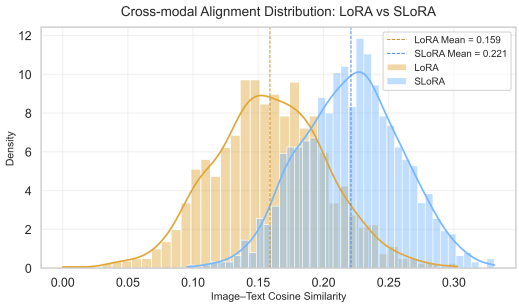

Figure 2: Cross-modal image-text similarity distribution on SLAKE using CLIP after LoRA and SLoRA adaptation.

static transformations that do not adapt to input variations. As a result, they provide limited control over both the distribution and the strength of adaptation updates.

To address these limitations, we propose **SLoRA**, a structured and adaptive extension of LoRA for cross-domain vision and vision-language transfer. SLoRA introduces two complementary mechanisms within the low-rank adaptation space. First, it performs a structured basis-space reparameterization and applies learnable basis scaling to redistribute adaptation capacity across different directions. Second, it employs an input-adaptive gating mechanism that dynamically controls the magnitude of LoRA updates according to the input characteristics. Together, these designs enable more flexible and stable adaptation while preserving the parameter efficiency and plug-and-play nature of LoRA.

We consider a practical PEFT setting where a pre-trained model is adapted to a labeled target domain while keeping the backbone frozen, without accessing source-domain samples. Extensive experiments on six domain-specific visual question answering (VQA) benchmarks spanning radiology, pathology, art, and remote sensing, together with six vision-language backbones, demonstrate the effectiveness of SLoRA. As shown in Figure 1, SLoRA consistently outperforms standard LoRA, achieving an average improvement of **2.2%** on VQA benchmarks, and in many cases matches or surpasses full fine-tuning while introducing only a moderate parameter overhead. Beyond multimodal VQA, SLoRA further achieves an average improvement of **2.9%** on eleven image classification datasets, confirming its robustness and generality across diverse domains.

We further analyze the effect of SLoRA on cross-modal alignment by visualizing the image-text cosine similarity distribution of a CLIP encoder adapted on the SLAKE Liu et al. (2021b) dataset. As illustrated in Figure 2, SLoRA yields a more compact similarity distribution with higher alignment scores, indicating improved and more stable image-text coupling under domain shift. These results suggest that structured low-rank modulation provides an effective and practical solution for robust parameter-efficient transfer across diverse models, tasks, and domains. Due to space constraints, additional related work, detailed theoretical analysis, algorithmic pseudocode, experimental details, ablation studies, and supplementary discussions are provided in the **Appendix**.

2 METHOD

Most PEFT methods, such as Adapters, Prefix-Tuning, and Low-Rank Adaptation (LoRA), freeze the pre-trained backbone and perform task adaptation via a low-rank update:

$$\Delta W = BA,$$

as illustrated in Figure 3. This formulation restricts optimization to a low-dimensional subspace, significantly reducing the number of trainable parameters while preserving pre-trained knowledge.

Despite its effectiveness, standard LoRA applies uniform adaptation across all low-rank directions. Under domain shifts, different directions in the low-rank space often contribute differently to adaptation: some directions may capture transferable representations, whereas others are more sensitive

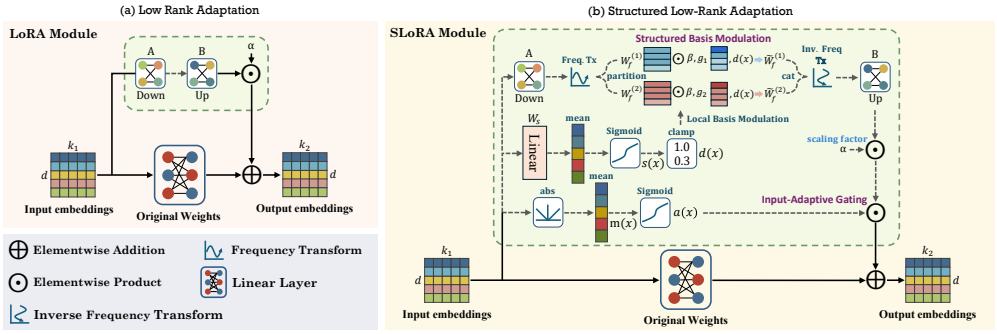

Figure 3: LoRA vs. SLoRA. SLoRA introduces structured basis modulation and input-adaptive gating on top of LoRA, enabling adaptive low-rank updates across heterogeneous adaptation directions.

to domain-specific variations. Treating all directions equally may therefore limit the flexibility of adaptation. To address this limitation, we introduce a structured reparameterization of the low-rank adaptation space. Instead of uniformly parameterizing the low-rank space, we introduce an intermediate structured representation in which the low-rank basis is expressed in a transformed coordinate system. This allows us to selectively modulate different subsets of basis directions in a structured manner while keeping the parameter budget unchanged.

2.1 Structured Basis Modulation

Low-rank Decomposition. Following LoRA, we parameterize the adaptation matrix as:

$$\Delta W = W_{up} W_{down},$$

where:

$$W_{down} \in \mathbb{R}^{r \times d_{in}}, \quad W_{up} \in \mathbb{R}^{d_{out} \times r}.$$

Only these parameters are updated during fine-tuning.

Orthogonal Basis Transformation. We apply a real-valued Fast Fourier Transform (FFT) along the rank dimension:

$$W_f = \mathcal{F}(W_{down}, \dim = 0), \quad W_f \in \mathbb{C}^{(r/2+1) \times d_{in}}.$$

This transformation reparameterizes the low-rank space into a fixed orthogonal basis. Since $\mathcal{F}(\cdot)$ is parameter-free, it does not increase model capacity but changes the coordinate system in which adaptation is expressed. FFT provides a fixed orthogonal-basis transformation with negligible computational overhead and no additional trainable parameters. This allows correlated rank directions to be represented as orthogonal basis coefficients in the transformed space, enabling structured modulation without additional parameters.

Basis Group Partition. In the transformed space, each row of W_f corresponds to a basis coefficient vector. We partition these coefficients into two groups according to a predefined partition ratio ρ:

$$W_f = [W_f^{(1)}, W_f^{(2)}],$$

where:

$$W_f^{(1)} \in \mathbb{C}^{k_1 \times d_{in}}, \quad W_f^{(2)} \in \mathbb{C}^{k_2 \times d_{in}}, \quad k_1 + k_2 = r/2 + 1.$$

This partition divides the transformed low-rank space into two structured subspaces, enabling independent control over different subsets of basis directions. Although more groups are possible, we empirically find that two groups provide a favorable trade-off between modeling flexibility, computational overhead, and optimization stability. We follow the real FFT formulation, where only the non-redundant transformed basis coefficients are retained, resulting in $(r/2 + 1)$ basis coefficients.

Learnable Basis Scaling. We introduce learnable modulation vectors:

$$g_1 \in \mathbb{R}^{k_1}, \quad g_2 \in \mathbb{R}^{k_2},$$

where each element of g_1 and g_2 corresponds to a single basis direction in the transformed space, controlling its contribution strength. We further introduce a global balancing coefficient:

$$\beta = \sigma(\theta_\beta) \in (0, 1),$$

which determines how adaptation capacity is distributed between the two basis groups.

Local Basis Modulation. Given a layer-normalized input $x \in \mathbb{R}^{d_{in}}$, we compute a scalar descriptor:

$$s(x) = \sigma(W_s^\top x), \quad W_s \in \mathbb{R}^{d_{in}}.$$

This projection provides a compact summary of the input activations. To ensure numerical stability across different batches, we bound this signal as:

$$d(x) = \mathrm{clamp}(s(x), 0.3, 1.0).$$

In practice, $d(x)$ is averaged across the mini-batch to reduce variance during training. We distinguish two complementary roles: $d(x)$ controls fine-grained modulation within the structured basis-space, affecting how adaptation capacity is distributed, while $a(x)$ controls the global magnitude of the LoRA update, determining how strongly adaptation is applied.

Modulated Basis Representation. We apply multiplicative modulation to each basis group:

$$\tilde{W}_f^{(1)} = W_f^{(1)} \odot \left(1 + \beta \tanh(g_1)\, d(x)\right),$$

$$\tilde{W}_f^{(2)} = W_f^{(2)} \odot \left(1 + (1-\beta) \tanh(g_2)\, d(x)\right),$$

where $\odot$ denotes row-wise broadcasting multiplication. Each basis coefficient is therefore scaled according to: (i) its group assignment, (ii) its learnable importance, and (iii) the local basis modulation. The full transformed representation is:

$$\tilde{W}_f = [\tilde{W}_f^{(1)}, \tilde{W}_f^{(2)}].$$

Inverse Transformation. The adapted low-rank projection is reconstructed via inverse FFT:

$$\tilde{W}_{down} = \mathcal{F}^{-1}(\tilde{W}_f).$$

This step maps the modulated basis representation back to the original low-rank parameter space, producing a task-adapted projection matrix. The modulation is applied on a structured basis-space while maintaining compatibility with the original LoRA parameterization, allowing seamless integration without modifying downstream computation.

2.2 Input-Adaptive Gating

In addition to basis modulation, we introduce a global gating mechanism that controls the overall strength of the LoRA update.

We define an activation magnitude descriptor:

$$m(x) = \frac{1}{d_{in}} \sum_i |x_i|.$$

This quantity measures the average magnitude of the input activations, providing a robust estimate of input intensity that is insensitive to sign cancellation. The gating function is defined as:

$$a(x) = \sigma(m(x)).$$

This scalar is applied uniformly to the LoRA update, ensuring that inputs with stronger activation magnitudes are allowed to produce larger adaptation updates.

Final Output. We retain the standard LoRA scaling (α/r), which bounds the magnitude of modulated updates. The final output is given by:

$$y = y_{\text{orig}} + a(x)\, W_{up}(\tilde{W}_{down}x),$$

where $a(x)$ controls the overall update magnitude, and $\tilde{W}_{down}$ encodes structured basis-modulation.

2.3 Summary of Computation Flow

Since LoRA lacks an explicit mechanism to organize adaptation capacity across low-rank directions, we employ a fixed orthogonal basis-space transformation that exposes structured basis directions. Given input x, SLoRA proceeds as follows: (1) compute an input-dependent descriptor to obtain the local modulation factor $d(x)$ and the global update gate $a(x)$; (2) transform LoRA parameters into a fixed orthogonal basis via FFT; (3) apply structured group-wise and element-wise modulation in the transformed space; (4) reconstruct the adapted projection via inverse FFT; (5) apply the globally gated LoRA update to produce the final output.

3 Experiments

We evaluate SLoRA on diverse cross-domain vision-language and vision benchmarks to comprehensively assess its effectiveness, robustness, and generalizability across different domains, tasks, and backbone architectures.

3.1 Experimental Setup

Primary Evaluation: Cross-Domain VQA. We adopt six VQA benchmarks as our primary evaluation setting, selected because they exhibit substantial domain shifts and diverse semantic reasoning requirements. Radiology datasets include VQA-RAD Lau et al. (2018) and SLAKE Liu et al. (2021b); pathology adaptation is assessed on PathVQA He et al. (2020); orthopedics is represented by OVQA Huang et al. (2022); and remote sensing and art domains are evaluated using EarthVQA Wang et al. (2024) and SemArt Garcia et al. (2020), respectively.

Supplementary Validation: Image Classification. We additionally evaluate SLoRA on eleven image classification benchmarks spanning the same four application domains. In the radiology domain, we adopt five MedMNIST datasets Yang et al. (2021), including OrganA, OrganC, OrganS, PneumoniaMNIST, and BreastMNIST. The pathology domain includes TissueMNIST and PathMNIST; remote sensing uses UCMerced Yang & Newsam (2010), AID Xia et al. (2017), and NWPU Cheng et al. (2017) with standard train/test splits (70%, 50%, and 70%); and the art domain is represented by AHE Llamas et al. (2017).

Pre-trained Backbones. We validate SLoRA across a diverse set of pre-trained architectures. For VQA, we use ViLT Kim et al. (2021), CLIP Radford et al. (2021), BLIP Li et al. (2022b), CoCa Yu et al. (2022), GIT Wang et al. (2022), and UForm unum-cloud (2023). For image classification, we evaluate using ViT Dosovitskiy et al. (2020), MAE He et al. (2022), CLIP, BLIP, BEiT Bao et al. (2021), and DiT Li et al. (2022a). Due to hardware constraints, all evaluated backbones contain fewer than 1.5B parameters.

Implementation Details. Unless otherwise stated, all experiments use a fixed LoRA rank $r = 16$ and scaling factor $\alpha = 32$ for all inserted layers across both image and VQA models, without model-specific, layer-specific, or dataset-specific tuning. For fair comparison, all LoRA-based baselines are applied to the exact same set of projection layers with identical rank and scaling configurations.

Training Protocol. All models are trained using four NVIDIA V100 GPUs (32GB) and optimized with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) using a base learning rate of 1×10^{-4}. The batch size ranges from 16 to 256 depending on the backbone capacity. RandAugment is applied to image data augmentation. Accuracy (Acc) serves as the primary evaluation metric for all tasks. SLoRA modules are inserted into the same projection layers as LoRA, while all other weights are frozen unless otherwise specified.

Table 1: Comparison with representative methods on radiology, pathology, and orthopedic VQA benchmarks. Baseline models* use LoRA fine-tuning, and † denotes the corresponding SLoRA variants. Model sizes are shown in blue for LoRA and red for SLoRA.

Methods	Radiology Domain						Pathology Domain			Cross Orthopedic		
	VQA-RAD			SLAKE			PathVQA			OVQA		
	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall
MMQ	53.7	75.8	67.0	-	-	-	13.4	84.0	48.8	56.9	76.2	68.5
MEVF	49.2	77.2	66.1	77.8	79.8	78.6	8.1	81.4	44.8	36.3	76.3	60.4
CR	60.0	79.3	71.6	77.8	79.8	78.6	-	-	-	51.6	77.7	67.7
CPRD	52.5	77.9	67.8	79.5	83.4	81.1	-	-	-			
MMBert	63.1	77.9	72.0	-	-	-	-	-	-	37.9	80.2	63.3
Med-CLIP	60.1	80.0	72.1	78.4	82.5	80.1	-	-	-	-	-	-
M3AE	**67.2**	83.4	77.0	80.3	87.8	83.2	-	-	-	-	-	-
PMC-CLIP	67.0	**84.0**	**77.6**	**81.9**	**88.0**	**84.3**	-	-	-	52.6	82.3	70.5
ViLT* (118.6M)	38.0	68.7	56.5	76.6	84.2	79.6	28.2	85.8	57.1	37.8	79.8	63.0
ViLT† (118.7M)	40.2	69.8	58.0	78.4	85.8	81.3	30.0	86.0	58.1	40.2	81.3	64.9
CLIP* (153.3M)	55.8	74.7	67.1	78.5	85.5	81.2	32.0	86.6	59.4	48.0	81.4	68.1
CLIP† (153.4M)	56.5	77.6	69.2	80.5	**88.0**	83.5	34.0	87.7	60.9	52.7	82.3	70.5
BLIP* (390.1M)	56.4	75.4	67.8	78.5	85.9	81.4	33.1	86.6	59.8	48.6	80.6	67.8
BLIP† (390.4M)	56.6	78.7	69.9	80.2	87.3	83.0	**35.3**	88.0	**61.7**	50.3	83.1	70.0
CoCa* (259.4M)	38.6	64.7	54.4	76.8	86.1	80.4	27.3	85.3	56.4	43.8	77.3	64.0
CoCa† (259.5M)	37.5	69.5	56.8	79.4	85.9	81.9	29.0	85.7	57.4	41.6	82.3	66.1
GIT* (155.5M)	60.9	78.6	71.6	79.7	85.6	82.0	31.0	86.8	59.0	52.5	82.4	70.5
GIT† (155.6M)	63.7	79.1	73.0	80.5	87.3	83.2	34.8	87.9	61.4	**59.3**	**84.6**	**74.5**
UForm* (1507.9M)	57.0	74.3	67.4	79.1	84.7	81.3	30.9	87.4	59.2	48.6	80.6	67.8
UForm† (1509.6M)	60.9	77.6	71.0	79.9	87.3	82.8	33.7	**88.1**	61.0	55.0	82.2	71.3

Table 2: Comparison with existing methods on VQA tasks across remote sensing and art.

Method	Remote	Art
	EarthVQA	SemArt
SAN	75.7	-
BAN	76.7	22.4
BUTD	76.4	21.8
Instruct-BLIP	75.2	-
RSVQA	70.8	-
SOBA	**78.1**	-
VIKING	-	20.4
ViLT*	72.6	26.5
ViLT†	74.6	30.1
CLIP*	74.9	28.4
CLIP†	76.4	30.7
BLIP*	74.4	28.3
BLIP†	76.8	30.9
CoCa*	71.2	25.4
CoCa†	73.1	28.9
GIT*	74.7	29.2
GIT†	77.0	**31.0**
UForm*	74.0	26.5
UForm†	76.4	29.0

3.2 Baselines

We compare SLoRA with both task-specific state-of-the-art methods and standard LoRA. For radiology and pathology VQA, we include models that are specifically designed or pre-trained with large-scale medical imaging corpora before fine-tuning on downstream VQA tasks, such as MMQ Do et al. (2021), MEVF Nguyen et al. (2019), CR Zhan et al. (2020), CPRD Liu et al. (2021a), MMBert Khare et al. (2021), Med-CLIP Eslami et al. (2021), M3AE Chen et al. (2022), and PMC-CLIP Lin et al. (2023). For remote sensing and art VQA, we report competitive methods widely used in those respective communities, including SAN Yang et al. (2016), BAN Kim et al. (2018), BUTD Anderson et al. (2018), Instruct-BLIP Dai et al. (2023), RSVQA Zheng et al. (2021), SOBA Wang et al. (2024), and VIKING Garcia et al. (2020).

For image classification, we adopt representative architectures from both CNN and transformer families: ResNet-18, ResNet-50, DeiT-B Touvron et al. (2021), ViT-B/L Dosovitskiy et al. (2020), Swin-B, VMamba-T Zhu et al. (2024), and also compare against AutoML frameworks (Auto-sklearn, AutoKeras, and Google ML) following the official MedMNIST evaluation protocol. Importantly, all compared methods are representative baselines for their respective domains and datasets.

3.3 Results Analysis

Multimodal VQA Results. To evaluate the effectiveness of SLoRA for cross-domain vision-language transfer, we compare it against standard LoRA across a comprehensive suite of VQA datasets spanning radiology, pathology, remote sensing, and art. Table 1 and Table 2 report results for six backbone architectures, including ViLT, CLIP, BLIP, CoCa, GIT, and UForm. Across all evaluated datasets and backbone architectures, SLoRA consistently outperforms the LoRA baselines, achieving an average improvement of **2.2%** in overall accuracy. These results indicate that structured low-rank modulation provides a more flexible and effective adaptation mechanism than standard LoRA, yielding consistent improvements across diverse vision-language tasks and domains.

Image Classification Results. To assess whether SLoRA's benefits extend beyond vision-language transfer, we further evaluate it on a diverse set of single-modal image classification benchmarks covering radiology, pathology, remote sensing, and art, as shown in Table 3 and Table 4. Unlike multimodal VQA, which requires cross-modal reasoning between visual and textual representations, image classification provides a complementary unimodal setting for evaluating whether the proposed structured low-rank adaptation generalizes beyond multimodal transfer. Across all six backbone architectures, SLoRA consistently outperforms standard LoRA, achieving an average accuracy improvement of **2.9%** across all datasets. Importantly, SLoRA also surpasses full-parameter fine-tuning in the majority of classification tasks (Appendix Table 22 and Table 23).

Training Dynamics. We further analyze the optimization behavior of SLoRA by comparing its training trajectory with standard LoRA on a multimodal task (SLAKE VQA) and a single-modal task (OrganA classification). Figure 4 shows the first 300 epochs of training loss and accuracy curves,

Table 3: Comparison with existing methods on image classification datasets from radiology and pathology domains.

Model	Radiology Domain					Pathology Domain	
	OrganA	OrganC	OrganS	Pneumonia	Breast	Tissue	Path
ResNet-18	0.935	0.900	0.782	0.854	0.863	0.676	0.907
ResNet-50	0.935	0.905	0.770	0.854	0.812	0.680	0.911
Auto-sklearn	0.762	0.829	0.672	0.855	0.803	0.532	0.716
AutoKeras	0.905	0.879	0.813	0.878	0.831	0.703	0.834
Google ML	0.886	0.877	0.749	0.946	0.861	0.673	0.728
ViT* (87.4M)	0.930	0.915	0.792	0.927	0.912	0.682	0.925
ViT† (86.3M)	0.963	0.941	0.828	**0.960**	**0.934**	0.713	0.954
MAE* (86.7M)	0.928	0.895	0.786	0.908	0.892	0.651	0.898
MAE† (86.7M)	0.942	0.927	0.809	0.932	0.915	0.668	0.934
CLIP* (118.8M)	0.912	0.894	0.792	0.917	0.894	0.697	0.893
CLIP† (118.9M)	0.958	0.935	0.831	0.940	0.915	0.706	0.942
BLIP* (388.8M)	0.941	0.874	0.767	0.901	0.922	0.688	0.915
BLIP† (389.1M)	0.955	0.928	0.809	0.959	0.932	0.697	0.946
BEiT* (103.4M)	0.915	0.911	0.791	0.903	0.905	0.693	0.925
BEiT† (103.5M)	**0.968**	**0.942**	**0.844**	0.934	0.917	**0.725**	**0.956**
DiT* (86.7M)	0.931	0.909	0.783	0.921	0.878	0.673	0.875
DiT† (86.8M)	0.952	0.932	0.826	0.950	0.899	0.695	0.894

Table 4: Comparison with existing methods on image classification tasks in remote sensing and art domains.

Model	Remote Sensing Domain			Art Domain
	UCMerced(70%)	AID(50%)	NWPU(70%)	AHE
ResNet-50	0.920	0.895	0.927	0.930
DeiT-B	0.891	0.874	0.891	0.934
ViT-B	0.911	0.894	0.889	0.941
ViT-L	0.920	0.902	0.913	0.946
Swin-B	0.919	0.899	0.937	0.940
VMamba-T	0.932	0.916	0.940	0.948
ViT*	0.965	0.933	0.941	0.948
ViT†	**0.991**	**0.960**	0.960	**0.966**
MAE*	0.936	0.901	0.904	0.901
MAE†	0.971	0.928	0.932	0.929
CLIP*	0.962	0.905	0.908	0.928
CLIP†	0.983	0.926	0.947	0.945
BLIP*	0.959	0.879	0.908	0.913
BLIP†	0.979	0.910	0.943	0.939
BEiT*	0.965	0.922	0.939	0.927
BEiT†	0.989	0.958	**0.962**	0.963
DiT*	0.925	0.863	0.859	0.921
DiT†	0.950	0.888	0.882	0.963

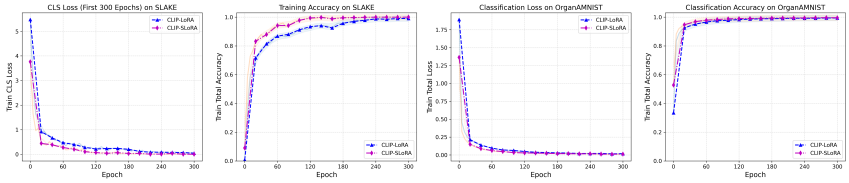

Figure 4: Training dynamics comparison between LoRA and SLoRA across different tasks.

with mild smoothing applied for clarity. Across both tasks, SLoRA consistently reduces training loss faster and reaches higher validation accuracy earlier during training. These results suggest that structured basis modulation and adaptive update control improve optimization efficiency, enabling more stable adaptation.

Across multimodal VQA, image classification, and optimization analyses, SLoRA consistently improves over LoRA across diverse architectures and domains. Taken together, these results suggest that restructuring the low-rank adaptation space through basis-space reparameterization, together with adaptive modulation and gating, leads to more effective transfer under domain shift.

Computational Efficiency and Overhead. We evaluate the training efficiency and computational overhead of SLoRA in comparison with standard LoRA and FouRA Borse et al. (2024) on CLIP-ViT using a batch size of 32 on a single NVIDIA V100 GPU. As shown in Table 5, SLoRA incurs a moderate increase in per-step training time (139.7 ms $\rightarrow$ 158.3 ms, +13.3%) and peak GPU memory usage (1.56 GB $\rightarrow$ 1.73 GB) relative to LoRA. This overhead arises from the additional basis-space modulation and adaptive gating operations applied to the low-rank adapters, while the FFT is applied only along the low-rank dimension ($r = 16$ in our default setting), making the additional computation negligible relative to backbone inference. Considering the consistent performance improvements across a wide range of tasks and architectures, this additional computational cost represents a favorable trade-off between efficiency and adaptation quality. Importantly, the reported step time is measured over the full training iteration, including forward, backward, and optimizer updates, ensuring a fair comparison.

3.4 Comparison with PEFT Methods

Table 6 compares SLoRA with representative PEFT methods, including LoRA variants (DoRA Liu et al. (2024b), CoLA Zhou et al. (2025)), adaptive-rank approaches (AdaLoRA Zhang et al. (2023), GaRA Lee et al. (2025)), and adaptation-space transformation methods (Spectral Adapter Zhang & Pilanci (2024), FourierFT Gao et al. (2024), FouRA Borse et al. (2024), and LoCA Du et al. (2025)). All methods are evaluated using the official implementations and recommended configurations while maintaining comparable parameter budgets whenever possible. Across all three backbones and datasets, SLoRA consistently achieves the best performance. Notably, SLoRA also outperforms recent transformed adaptation methods such as FouRA and LoCA, despite introducing only marginal parameter overhead. These results suggest that the performance gains arise not

Table 5: Training efficiency and computational overhead of LoRA, FouRA, and SLoRA on CLIP-ViT. Relative step-time overhead is reported with respect to standard LoRA.

Method	Params	Train params	Step Time (ms)	Memory (GB)	FFT Overhead
LoRA	153.3M	0.60M	139.69	1.56 G	-
FouRA	158.5M	0.84M	159.26	1.70 G	+14.0%
SLoRA	153.4M	0.71M	158.30	1.73 G	+13.3%

Table 6: Comparison of representative PEFT methods across LoRA variants.

Method	CLIP				BLIP				GIT			
	#Params	VQA-RAD	PathVQA	OVQA	#Params	VQA-RAD	PathVQA	OVQA	#Params	VQA-RAD	PathVQA	OVQA
Fine-tuned	-	66.7	58.1	70.1	-	66.5	58.0	67.1	-	70.1	58.1	69.9
LoRA	1.4%	67.1	59.4	68.1	1.0%	67.8	59.8	67.8	1.2%	71.6	59.0	70.5
DoRA	1.7%	68.0	60.2	69.2	1.2%	69.0	61.1	69.1	1.4%	72.0	60.6	72.9
Spectral	1.4%	68.7	60.4	69.8	1.0%	69.1	60.8	69.0	1.2%	72.2	61.0	73.2
FourierFT	0.4%	67.3	59.4	68.0	0.5%	68.0	59.9	68.0	0.5%	71.9	59.1	70.3
CoLA	1.4%	68.0	59.8	68.9	1.1%	68.8	60.4	68.7	1.3%	72.4	61.0	72.4
FouRA	1.6%	68.2	60.0	69.3	1.1%	69.0	60.2	69.3	1.3%	71.4	60.1	73.1
AdaLoRA	1.5%	68.0	59.9	68.6	1.1%	68.6	59.9	68.3	1.3%	71.4	60.4	72.6
GaRA	1.5%	68.4	60.0	69.2	1.2%	69.1	60.6	69.1	1.4%	72.0	60.6	73.1
LoCA	1.6%	68.6	60.3	69.2	1.2%	68.6	61.0	69.4	1.4%	72.5	60.7	73.2
SLoRA	1.4%	**69.2**	**60.9**	**70.5**	1.0%	**69.9**	**61.7**	**70.0**	1.2%	**73.0**	**61.4**	**74.5**

merely from transforming the adaptation space or increasing expressive capacity, but from jointly combining structured basis reparameterization with input-adaptive update control. Simply increasing adaptation capacity or applying fixed spectral transformations is insufficient. Instead, effective PEFT under domain shift requires jointly controlling the structure of low-rank updates and their local basis modulation. By integrating basis reparameterization with adaptive gating in a unified framework, SLoRA achieves more stable and transferable improvements. This design improves adaptation across architectures, datasets, and domains.

3.5 Ablation Study on Rank and Scaling

Table 7 investigates the influence of the LoRA rank r and scaling factor α. Increasing either parameter does not consistently lead to better performance. Although larger values increase the adaptation capacity, they may also introduce unnecessary flexibility that weakens optimization stability. Across all three backbones, the configuration $(r = 16, \alpha = 32)$ achieves the most consistent results. Smaller ranks limit adaptation capacity, whereas excessively large settings provide little additional benefit and occasionally degrade performance. These observations indicate that SLoRA benefits from structured adaptation rather than simply increasing parameterization.

3.6 Ablation Study on SLoRA Key Components

To understand the contribution of each component, we progressively enable basis-space transformation, structured basis modulation, and input-adaptive gating. As shown in Table 8, basis-space transformation alone provides limited gains, indicating that reparameterization itself is insufficient for effective adaptation. Adding structured basis modulation or input-adaptive gating further improves performance by enhancing basis-level adaptation and adaptive update control, respectively. The best performance is achieved when all components are combined. Figure 5 further shows that the complete SLoRA model converges faster and exhibits smoother optimization dynamics than its partial variants. These results demonstrate that the three components are complementary and jointly contribute to both adaptation quality and training stability.

3.7 Additional Experiments

Due to space constraints, several complementary analyses are provided in the **Appendix**. These include detailed implementation configurations, optimization settings, complete experimental results, and extensive ablation studies covering random-seed robustness, nonlinear activation functions, input-adaptive gating, structured basis modulation, and basis-space reparameterization. We further compare different basis-space reparameterization strategies and evaluate SLoRA on recent multimodal large language models to verify its generalization beyond conventional vision-language

Table 7: Ablation study of SLoRA with varying rank (r) and scaling factor (α) on three VQA benchmarks.

Model	r	α	VQA-RAD	PathVQA	OVQA
CLIP	4	16	63.9	60.3	69.0
	4	32	63.2	60.3	69.2
	8	32	67.5	60.4	68.9
	16	16	66.1	60.5	68.2
	16	32	**69.2**	**60.9**	**70.5**
	16	64	66.8	60.8	**70.5**
	32	64	65.5	60.0	70.1
BLIP	4	16	68.9	60.3	68.3
	4	32	67.0	60.7	69.1
	8	32	67.5	60.7	69.2
	16	16	69.5	60.6	69.1
	16	32	**69.9**	61.7	**70.0**
	16	64	67.7	61.2	69.2
	32	64	67.0	**61.8**	68.9
GIT	4	16	70.5	59.9	71.0
	4	32	70.5	60.5	72.0
	8	32	70.3	60.2	72.6
	16	16	71.8	60.8	72.8
	16	32	**73.0**	**61.4**	**74.5**
	16	64	71.6	61.2	72.4
	32	64	72.5	**61.4**	73.6

Table 8: Ablation study of the key components in SLoRA.

Model	Basis Space Transform	Structured Basis Modulation	Input Adaptive Gating	VQA-RAD	PathVQA	OVQA
CLIP				67.1	59.4	68.1
	✓			67.1	59.7	68.8
	✓		✓	67.8	59.9	69.1
	✓	✓		68.4	60.4	69.7
			✓	67.1	59.9	68.8
	✓	✓	✓	**69.2**	**60.9**	**70.5**
BLIP				67.8	59.8	67.8
	✓			67.8	60.3	68.3
	✓		✓	68.2	60.8	68.8
	✓	✓		69.0	61.2	69.6
			✓	68.0	60.8	68.4
	✓	✓	✓	**69.9**	**61.7**	**70.0**
GIT				71.6	59.0	70.5
	✓			71.6	59.4	71.4
	✓		✓	72.1	60.6	73.0
	✓	✓		72.4	60.8	74.2
			✓	72.1	60.6	72.5
	✓	✓	✓	**73.0**	**61.4**	**74.5**

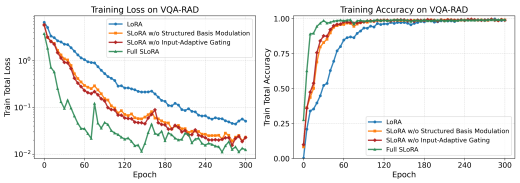

Figure 5: Training dynamics comparison between LoRA and SLoRA variants on VQA-RAD using CLIP. Left: training total loss (log scale). Right: training total accuracy.

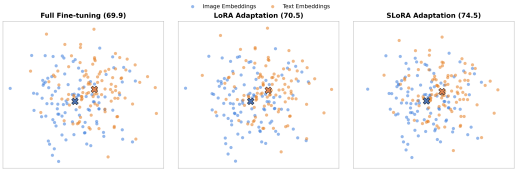

Figure 6: t-SNE visualization of the GIT model embeddings on the OVQA dataset under different adaptation methods.

backbones. Collectively, these results demonstrate the stability, effectiveness, and broad applicability of SLoRA across diverse architectures and training conditions.

4 Feature Alignment Analysis

Figure 6 visualizes image-text embeddings of GIT on OVQA using t-SNE. Compared with full fine-tuning and standard LoRA, SLoRA produces more compact feature clusters and a smaller discrepancy between image and text centroids. The tighter alignment suggests that structured basis modulation promotes more consistent cross-modal representations under domain shift, which correlates with the substantial performance gains observed on OVQA. This improved alignment also reduces cross-modal ambiguity under domain shift, leading to more reliable reasoning and improved performance. These observations provide qualitative evidence that SLoRA improves adaptation not only at the prediction level but also in the underlying representation space.

5 Conclusion

In this work, we presented SLoRA, a structured low-rank adaptation framework for parameter-efficient transfer under domain shift. Rather than directly optimizing low-rank updates in the original parameter space, SLoRA introduces a lightweight basis-space reparameterization together with structured basis modulation and input-adaptive gating, enabling more expressive yet controlled adaptation without substantially increasing parameter complexity. Extensive experiments across six cross-domain VQA benchmarks, eleven image classification datasets, and multiple vision-language backbones demonstrate that SLoRA outperforms standard LoRA and a wide range of representative PEFT methods. The results further show that structured adaptation can offer a favorable balance between transferability, robustness, and efficiency, making SLoRA a practical and effective PEFT solution for domain-shifted vision and vision-language tasks.

AI USE STATEMENT

Generative AI tools were used solely for language editing and polishing to improve the clarity, grammar, and readability of the manuscript. They were not used to generate research ideas, develop the proposed method, design or conduct experiments, analyze experimental results, or formulate scientific claims. All AI-assisted edits were carefully reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed SLoRA method, model configuration, datasets, evaluation protocols, and experimental settings in the main paper and appendix. Additional implementation details, hyperparameter settings, ablation studies, and supplementary experimental results are also provided to facilitate reproduction of the reported results.

REFERENCES

Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Shubhankar Borse, Shreya Kadambi, Nilesh P Pandey, Kartikeya Bhardwaj, Viswanath Ganapathy, Sweta Priyadarshi, Risheek Garrepalli, Rafael Esteves, Munawar Hayat, and Fatih Porikli. Foura: Fourier low-rank adaptation. *Advances in Neural Information Processing Systems*, 37:71504–71539, 2024.

E Oran Brigham. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pp. 679–689. Springer, 2022.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235, 2023.

Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pp. 64–74. Springer, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zhekai Du, Yinjie Min, Jingjing Li, Ke Lu, Changliang Zou, Liuhua Peng, Tingjin Chu, and Mingming Gong. Loca: Location-aware cosine adaptation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2502.06820*, 2025.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. Parameter-efficient fine-tuning with discrete fourier transform. *arXiv preprint arXiv:2405.03003*, 2024.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 92–108. Springer, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

A Hedayat and Walter Dennis Wallis. Hadamard matrices and their applications. *The annals of statistics*, pp. 1184–1238, 1978.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. Ovqa: A clinically generated visual question answering dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2924–2938, 2022.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1033–1036. IEEE, 2021.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Sohyun Lee, Yeho Gwon, Lukas Hoyer, and Suha Kwak. Gara-sam: Robustifying segment anything model with gated-rank adaptation. *arXiv preprint arXiv:2506.02882*, 2025.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3530–3539, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022b.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 210–220. Springer, 2021a.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654. IEEE, 2021b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Lei Liu, Xiangdong Su, and Guanglai Gao. Fourier self-adaptation for transferring general pre-trained models to specific domains. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3605–3614, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024b.

Jose Llamas, Pedro M. Lerones, Roberto Medina, Eduardo Zalama, and Jaime Gómez-García-Bermejo. Classification of architectural heritage images using deep learning techniques. *Applied Sciences*, 7(10):992, 2017.

Yuchen Mao, Hongwei Li, Wei Pang, Giorgos Papanastasiou, Guang Yang, and Chengjia Wang. Selora: Self-expanding low-rank adaptation of latent diffusion model for medical image synthesis. *arXiv preprint arXiv:2408.07196*, 2024.

Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 522–530. Springer, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

unum-cloud. Pocket-sized multimodal ai for content understanding and generation. `https://github.com/unum-cloud/uform`, 2023.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5481–5489, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2345–2354, 2020.

Fangzhao Zhang and Mert Pilanci. Spectral adapter: Fine-tuning in spectral space. *arXiv preprint arXiv:2405.13952*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Yiyun Zhou, Chang Yao, and Jingyuan Chen. Cola: Collaborative low-rank adaptation. *arXiv preprint arXiv:2505.15471*, 2025.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.